MASTER'S THESIS

A data analytics methodology for a fast-changing discipline

Ingen van, E (Erik)

Award date:
2020

A data analytics methodology for a fast-changing discipline

Degree programme:	Open University of the Netherlands Faculty of Management, Science & Technology Business Process Management & IT master's programme
Course:	IM9806 Master's Thesis BPMIT
Identification number:	
Student:	Erik van Ingen
Date:	June 28, 2020
Thesis supervisor:	Jeroen Baijens, MSc
Second reader:	Prof.dr.ir. Rob Kusters
Version number:	1.0
Status:	Final

Abstract

Massive investments are being made in Data Analytics (DA) but there is no substantial effort to mature the working methods. Around 82% of DA practitioners do not use any methodology in their projects, while 85% believe it would improve their work. Organizations that rely on ad hoc processes (as opposed to planned processes) are only half as likely to rate their projects as successful. This study helps DA practitioners to overcome this contradiction.

A design science research method was chosen in combination with case study research to design a new matrix of project and methodology classes and their connections. The matrix is strongly rooted in existing DA methodologies like CRISP-DM and Snail Shell with the difference that the matrix is *layered*. Fourteen DA experts were consulted in interviews and focus groups to reflect on a new matrix.

The research resulted in the design of project classes such as hypothesis generation/testing, big data, self-service analytics and data governance and related them to the process steps of the DA methodology. This helps DA practitioners to make an elaborated methodology choice while also being able to gear it to their preferred orientation in terms of agile, iterative or waterfall.

Key terms

data analytics methodology, agile, self-service analytics, cloud native, pipeline, continuous integration

Summary

Modern technologies allow for the generation and collection of (big) data. This data is used by companies to get more insights through data analytics projects. Practitioners in Data Analytics Projects (DAP) tend to use an ad hoc process, while they largely believe they would benefit from using a Data Analytics Methodology (DAM). Using an ad hoc process in data analytics leads to numerous problems and increases the risk of failure of DA projects. The problems relate to team efficiency, information sharing, delivering the 'wrong thing', no reproducibility, coordination and scope creep.

Data analytics (DA) in formal terms is the practice of descriptive, predictive and prescriptive analytics and includes business analytics, business intelligence, data warehousing and data science. DA creates sustainable value for business, creates insights for better decision making that leads to increased company performance. Massive investments are made in Data Analytics (DA) but there is no substantial effort to mature the working methods. Around 82% of DA practitioners do not use any DAM in their DAP, while 85% believes it would improve their work. This study helps DA practitioners to overcome this contradiction.

The main research question is: What instruments are needed to convince DA practitioners to use a DAM? Subsequent questions are: What are the DAP and DAM classes and how they be mapped?

A design science research method was chosen in combination with case study research to design a new matrix of DAP and DAM classes and their connections. The matrix is strongly rooted in existing DAMs like CRISP-DM and Snail Shell with the difference that the matrix is *layered*. Fourteen DA experts were consulted in interviews and focus groups to design the research artefact.

Various DAP classes are designed or discovered as hypothesis generation/testing, big data, self-service analytics and data governance and related to the process steps of the DAM. This helps DA practitioners to make an elaborated choice while being able to gear it to their preferred methodology in terms of agile, iterative or waterfall.

Projects involving hypothesis generation generally are better suited to an iterative or agile approach in order to be able to anticipate uncertainties and due to their (often) explorative character. A hypothesis-generation project tends to start with business understanding and then moves into the data process steps. A hypothesis-testing project starts at the first process step, the problem formulation.

In terms of waterfall or agile, this study discovered a missing concept called 'iterative'. The concept 'iterative' refers to those projects that work fully iteratively but have not adopted a formal agile methodology like SCRUM or Kanban and/or do no use agile practices like continuous integration.

The reader is invited to read the full report to get insights into which methodology to choose and to understand emerging concepts like self-service analytics and cloud native which are changing the DA as a discipline.

Table of Contents

Table of Figures

Table of Tables

1. Introduction

1.1. Background

Modern technologies allow for the generation and collection of (big) data. This data is used by companies to get more insights through data analytics projects. Practitioners in data analytics projects tend to use an ad hoc process, *while they largely believe they would benefit from using a formal process methodology*. Using an ad hoc process in data analytics leads to numerous problems which increase the risk of failure of DA projects. The problems relate to team efficiency, information sharing, delivering the 'wrong thing', reproducibility, coordination and scope creep (Spoelstra, J., Zhang, H., & Kumar, 2016). The objective of this research is to investigate the conceptual space of this contradiction and to help data analytics practitioners to break out of this vicious circle.

This chapter makes the reader familiar with the topic, context and objectives of the research. Chapter 2 covers the literature study and chapter 3 explains the methodology used for the research applied in chapters 4-6.

1.2. Exploration of the topic

Data analytics (DA) in formal terms is the practice of descriptive, predictive and prescriptive analytics. DA is often also referred to as Business Analytics (Lepenioti, Bousdekis, Apostolou, & Mentzas, 2020). Descriptive Analytics is about *what happened* and provides information through a data warehouse and reports. Predictive analytics is about *what will happen* and uses data mining algorithms and machine learning. Prescriptive analytics is about *what should happen* using simulation and decision modelling (Sharda et al., 2018). DA creates sustainable value for business. DA creates insights for better decision-making, which leads to increased company performance. (Wixom, Yen, & Relich, 2013).

In 2012, Harvard Business Review published the article "Data Scientist: The Sexiest Job of the 21st Century" (Davenport & Patil, 2012). This article highlighted the phenomena that was happening at that time (and never stopped!); the growing availability of enormous amounts of data (big data) and its usefulness for data analytics. Data Scientists enriched the field of DA with techniques like Machine Learning, a subfield of Artificial Intelligence, which is currently a hot topic.

DA is often practised in organisations through projects. These projects follow a process model or methodology. This research is about the methodologies used by DA practitioners. Discussing methodologies is often intertwined with discussing process. A process model defines what to do and a methodology defines how to do it (Mariscal, Marbán, & Fernández, 2010). This research complies with this definition and uses the term Data Analytics Methodologies (DAM). For projects, the term Data Analytics Projects (DAP) is used. *This study uses the word 'class' as a synonym of the word 'type'*. There are various classes to describe DAPs, examples are the above mentioned descriptive, predictive or prescriptive analytics (Sharda et al., 2018).

Within the field of DAM, there are formal methodologies like CRISP-DM and KDDS (Saltz et al., 2018). They provide a linear way to conduct a DA project. A different way to conduct these projects is with agile methodologies. Agile methodologies originate from the software engineering discipline and provide the organization with an iterative and flexible way of conducting a project. Example of these methodologies are SCRUM and Kanban (Ullah, 2019). More recent DAMs have been influenced by agile methodologies as discussed more in the following chapters.

1.3. Problem statement

Using a systematic process methodology anticipates on problems that teams face while using ad hoc processes. The problems could include slow information sharing, delivering the wrong thing, lack of reproducibility, inefficiencies and scope creep. Enormous investments are made in DA, but projects tend to rely on ad hoc methodologies. Around 82% of DA practitioners in projects does not use any DA methodology, while 85% of DA practitioners think that their data science efforts would improve if they used a systematic process methodology (Saltz, Hotz, Wild, & Stirling, 2018). These two numbers represent an apparent contradiction and the underlying causes are not very clear. This research aims to help DA practitioners to escape from this contradiction.

Massive investments are being made in DA but there is no substantial effort to mature the working methods, and to professionalize the discipline (Grover, Chiang, Liang, & Zhang, 2018). DAMs intend to mature the working methods and to professionalize the discipline. However, there is no research is available to help DA practitioners understand which DAM would fit well, given precise DA project characteristics. DA practitioners do not know which DAM to choose because they receive little guidance.

1.4. Research objective and questions

The objective of this research is to find the relevant classes of DAPs and DAMs, including their mappings. Examples of DAP classes come from the data science project model (Saltz, Shamshurin, & Connors, 2017), which identifies *infrastructure* (level of computing needs) and *discovery* (level of clarity of questions).

DAM classes take into consideration the existing DAMs like CRISP-DM or Snail Shell (*see also Appendix F CRISP-DM and Snail Shell)*. The DAM classes list/model also identifies possible new DAM classes that are needed to cover the conceptual space. A mapping indicates which DAM class most is most suitable for which DA project class; and helps practitioners to make a rational choice for a DAM.

The main research question is: *What instruments are needed to convince DA practitioners to use a DAM?* Two sub research questions are defined:

- RQ1: What are the DAP classes?
- RQ2: What are the DAM classes?

1.5. Motivation/relevance

Organizations that rely on ad hoc processes (as opposed to planned processes) are only half as likely to rate their projects as successful, when it comes to big data initiatives (Colas, Finck, Buvat, Nambiar, & Raj Singh, 2014). Therefore, moving away from ad hoc approaches is highly recommended. This research intends to help DA practitioners to make the move towards adopting a formal DAM.

This research contributes to decreasing waste of investments in DA projects. The research clarifies and enriches the conceptual space of DAMs, which is its scientific contribution. The research also provides guidance for the DA practitioners, which contributes to the professionalization of the DA discipline.

The software engineering community reports agile adoption rates varying from 14-40% (Abdalhamid & Mishra, 2017). Much of the DA community does *not* use a methodology (82%). This report helps to provide the necessary clarity to convince the DA community to adopt DAMs by explaining which DAM works best, given a certain context. It helps the reader to make an informed choice based on useful concepts found in the DAM scientific literature. Furthermore, the report intends to help the DA community to choose a methodology, thus hopefully increasing the adoption of methodologies.

1.6. Main lines of approach

The research is conducted according the principles of Design Science Research (DSR). DSR intends to design knowledge artefacts as a means to understanding the problem domain in a new and innovative way (Hevner, March, Park, & Ram, 2004). DSR is formalized by Peffers through six activities which can be performed in an iterative way (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007). One of the DSR activities is evaluation, which is further formalized by Venable in four steps (Venable, Pries-Heje, & Baskerville, 2016).

The next chapter provides the theoretical framework that includes the literature study. The literature study first clarifies the various types of DA projects. The literature study takes stock of the various DAMs used by DA practitioners. The result of the literature study is a theoretical overview that supports subsequent DSR iterations, developing a research artefact.

Within the container of the DSR methodology, case studies have been conducted in seven organizations to understand which DAMs are being used as of today. Semi-structured questions have been developed, building on concepts found in literature, with the intention of finding an answer to the research question(s). The DSR methodology has been used to discover the answers by developing so-called knowledge artefacts through the six DSR activities. This research clarifies whether a proper DAM is still lacking or whether existing DAMs just require polishing.

2. Theoretical framework

This section provides the theoretical framework.

2.1. Research approach

The purpose of this literature review is to understand the existing DAM body of knowledge. The literature review aims to discover the various DAP/DAM classes. Furthermore, the purpose of this literature review is to understand whether the research questions defined are indeed relevant or if they have already been covered by existing research.

The research approach for the theoretical framework uses a systematic literature review (Okoli & Schabram, 2010) in a number of steps. The steps involve describing the purpose of the literature review, how literature is searched for, quality appraisal, data extraction, synthesis and the review itself.

2.2. Implementation

The query was developed for searching articles on the field 'Title' (searching by topic returned too many articles). The query selects articles with the combination of two concepts in the title, DA and DAM. The query would also find this research if the main research question were published with the title: *What instruments are needed to convince data analytics practitioners to use a methodology?*

The starting point of the literature review are eight articles provided by the tutor of the Master Thesis course. This list of base articles is reflected in *Appendix A Base articles*. The only source used for finding articles was the Web of Science (http://webofknowledge.com/) from the company Clarivate Analytics. Only peer-reviewed research articles such as journal articles and conference proceedings are considered. This is the composed query to find the articles:

	key terms (AND)		
	data	methodologies	
related terms (OR)	datamining	project process	project methodology
	analytics	project method	process view
	big data	methodology	methodologies
	bigdata	methodological	process model
	data analytics	project	projects
	data science		

Table 1: Search terms

All the keywords in each cell of Table 1 are OR statements. There is only one AND statement which connects the two series of OR statements. The query uses the words 'project' and 'projects' for methodologies, while they are also part of other search terms. It would be expected that the other search terms are therefore redundant; however, this is not the case. This discrepancy is documented further in *Appendix B Query development*. It also specifies which studies were included for the review. The search query resulted in 222 results.

The practical screen (Okoli step 4) describes which articles were excluded and why. A total of 29 articles have been selected from the 222 based on the following criteria:

- The article should not discuss a domain specific DA case.
- The article needs to reflect on a combination of DA and DAM.

In most cases, only analysing the title suffices to understand whether the article meets the criteria. In some cases, the abstract of the article was read to make the final decision.

The quality appraisal (Okoli step 5) steps into the content of the article and establishes quality criteria. The abstracts of the 29 articles are evaluated against the same two above mentioned criteria and appraised on a scale from 0 to 10, along with the following additional quality criteria:

- The article must follow a sound scientific approach.
- The article needs to discuss DAMs, not IT tooling for supporting DAMs.
- The article needs to focus on DAMs, not only on its ethical aspects.

From the 29 articles, 8 articles scored 7 or higher. The 8 articles were analysed with these results:

- One article was not freely accessible (Mousannif, Sabah, Douiji, & Oulad Sayad, 2016).
- One document was not relevant because it was not about DAM but about designing a data warehouse (Tria, Lefons, & Tangorra, 2018)
- Three articles were already provided by the OU.
- Two documents were selected to incorporate further in this research: (Batra, 2018) and (Sharma, Osei-Bryson, Kasper, 2012).

The eight base articles plus two from the quality appraisal result in a total ten of articles. These ten articles were further analysed on research type and quality (see also *Appendix C Quality Appraisal)*. Three articles are based on qualitive research, three articles are based on quantitative research and two of them use a mix of both. All articles were found to have reasonable high-quality research approaches.

The data extraction (Okoli step 6) extracts all the applicable information from the selected articles (see *Appendix D Data Extraction* for the records). The articles (Gao, Koronios, & Selle, 2015; Saltz et al., 2018) had no explicit research model. The other six articles have research models and they have been extracted and listed in a separate document as a research record.

The synthesis of studies (Okoli step 7) is based on the ten articles. *Appendix E DAP and DAM classes* states for every article the concepts, definitions and its contribution to the body of knowledge.

2.3. Results and Conclusion

This section starts with looking into DAP classes, followed by looking into DAM classes.

Looking at DAP classes, some DAP classes relate directly to the project, others relate to the context of the project, e.g. the organization.

DAPs can be divided into those who work on **hypothesis** generation and those who work on hypothesis testing (Saltz et al., 2017). The work on hypothesis generation is more exploratory, whereas the work on hypothesis testing is much clearer and planned upfront. In case of hypothesis generation, the DA practitioner has almost carte blanche to find new knowledge in the data. This new knowledge can be found in the form of patterns or relations between one or more variables,

represented by the data. In the case of hypothesis testing, the patterns or relations are prescribed, and the DA's work is to quantify or qualify the relation to the best of their ability.

There are various types of **data**; it can be structured, unstructured, small and big data (Saltz et al., 2017). The aspects of big data are popularized and specified with the four V's of big data. The four V's of data are *Variety* (number of data sources), *Velocity* (speed at which the data changes), *Veracity* (trustworthiness) and *Volume* (size). However, the four V's do not suffice to describe all DA project as a whole (Saltz et al., 2017; Sharda et al., 2018). The 'big data' term applies both to structured and unstructured data. Big data projects tend to be large-scale where enormous size and heterogeneous data is transformed into structured data (Jensen, Nielsen, & Persson, 2019). This practice is more often associated with projects that do data science and predictive analytics. New DAMs like BAP are specifically geared towards big data projects (Gao et al., 2015). Instead, the more traditional descriptive analytics (business intelligence) projects are more often associated with *smaller data*. This tends to be structured data stored in a data warehouse which has been loaded from databases from within the organization.

The literature talks about **requirements** in very general terms like business-, stakeholder-, user-, project- and product requirements (Jensen et al., 2019) but they indicate only where the requirement comes from and do not indicate a class of requirements. However specific DA requirement classes are mentioned in problem descriptions which relate to team efficiency, information sharing, delivering the 'wrong thing', reproducibility, coordination and scope creep (Spoelstra, J., Zhang, H., & Kumar, 2016). These problems are here referred to as **requirement** classes. The requirements may differ a lot per DA project, as per instruction from senior management. It is also worth mentioning that it may be important for a project that all results are *reproducible* after project completion for scientific or data regulation reasons. In addition, a project may receive instructions that *scope creep* is to be avoided at all cost.

Org data governance in Figure 1 indicates a project context concept and it is referred to as *the organisational data governance*. Different publications use different names for this; such as knowledge management (Ahangama & Poo, 2015), information strategy (Gao et al., 2015) and enterprise data architecture (Li, Thomas, & Osei-Bryson, 2016). Lack of data governance is one of the biggest challenges in big data projects (Li et al., 2016). There is a need for a big data strategy as part of data governance in order to anticipate particular challenges related to the variety, velocity, veracity and volume of big data (Batra, 2018; Saltz et al., 2017). The lack of appropriate data governance leaves the control of data quality and integrity in the hands of application developers and research has not addressed this sufficiently (Li et al., 2016). The class *org data governance* is put forward to understand whether the data governance maturity of the organization affects the way that the methodology is used.

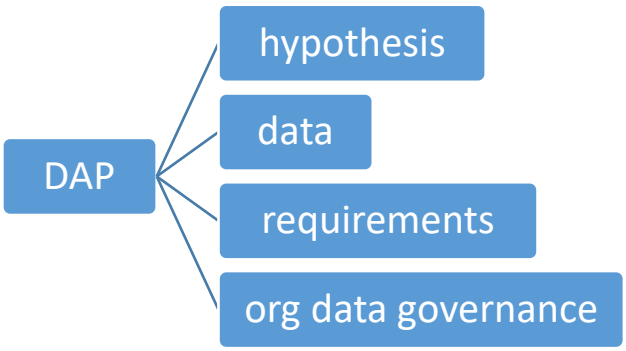


Figure 1: DAP dimension model

The DAP dimension model in Figure 1 illustrates the high-level classes found in the literature.

Looking at DAM classes, existing DAMs like CRISP-DM provide for useful classes. CRISP-DM is a phase based model where moving back and forth between phases is possible (Mariscal, Marbán, & Fernández, 2010). Many studies refer to CRISP-DM as widely known/used but state that CRISP-DM does not suffice anymore and present/propose new DAMs:

(Gao, Koronios, & Selle, 2015)	Business Analysis Process (BAP)
(Li, Thomas, & Osei-Bryson, 2016)	Snail Shell
(Mariscal, Marbán, & Fernández, 2010)	Refined Data Mining Process (RDMP)
(Sharma & Osei-Bryson, 2010)	Integrated Knowledge Discovery and Data Mining

Table 2: New presented models

Observing the DAMs, they are increasingly influenced by agile values (Batra, 2018). Batra groups DAMs into plan-driven, agile-plan balanced and agile-heavy. Plan- driven refers to a classical waterfall managed project, completely pre-planned. Agile-heavy methodologies are iterative in nature (such as BAP and Snail Shell). The distinction between plan driven and agile heavy is defined as **orientation** in the DAP-DAM matrix below. The class acknowledges the increasing influence of the agile paradigm on DA as opposed to waterfall. The orientation subclasses are mostly exclusive, even though it is possible for a project to apply a mix of waterfall and agile practices. A waterfall-oriented project performs the process steps only once and in the given order (from left to right). An agile-oriented project goes through the process steps (also in the given order) in multiple cycles. Examples of agile methodologies are SCRUM and Kanban. SCRUM divides a project into so-called sprints which range from one week to one month. Every sprint plans its sprint releasable deliverables, aligning continuously with the objectives and requirements of the stakeholders. Kanban is les prescriptive than SCRUM and works with the so-called Kanban board with 'TO DO', 'Doing' and 'Done'. Kanban aims to minimize the simultaneous pieces of work in progress by using limits to increase efficiency and results in data science have been promising (Saltz et al., 2018). The underlying idea is that having too many workstreams active at the same time leads to a loss of focus/productivity.

All models (regardless their *orientation*), talk about **process steps** or phases. Mariscal presents a new layered model with three high level process steps (**analysis, development and maintenance**). They are based on what Mariscal calls *phases* and sometimes *processes*. The *analyses* process step selects the DAM (which Mariscal refers to as the lifecycle model), it explores the problem domain, identifies the human resources needed for the project and performs data prospecting and data cleaning. The *development* process step pre-processes the data, and builds and improves the model through data reduction, data projection and data mining. The *maintenance* process step involves model updates, backups, data updates and software updates. (Mariscal et al., 2010). Considering the three models (CRISP-DM, BAP and Snail Shell), Snail Shell is the most complete model because it adds the notion of the *problem formulation*. The Snail Shell model is agile oriented, it is fairly recent (2016), fairly well cited. The Snail Shell classes (Li et al., 2016) are mapped onto the Mariscal phases to structure the classes from the Snail Shell model. Within those three high-level steps, the process steps from Snail Shell can be grouped. The analysis process step consists of the *problem formulation*, business and data understanding. The *development* process step involves the *data preparation* and *modelling*. The maintenance process step involves evaluation and deployment of the model. Snail Shell explicitly distinguishes between *problem formulation* and *maintenance*. The *problem formulation* step is useful because it forces us to think through why certain DA work is initiated. This is different

from the *business understanding*, which relates to the context of where the problem resides. The *maintenance* step is useful because, being a data project, data grows continuously and this tends to affect the model over time, which then needs *maintenance*. The DAM process steps are interdependent and each one of them helps to drive the subsequent process steps. Ignoring one step can result in problems in succeeding process steps (Sharma et al., 2012).

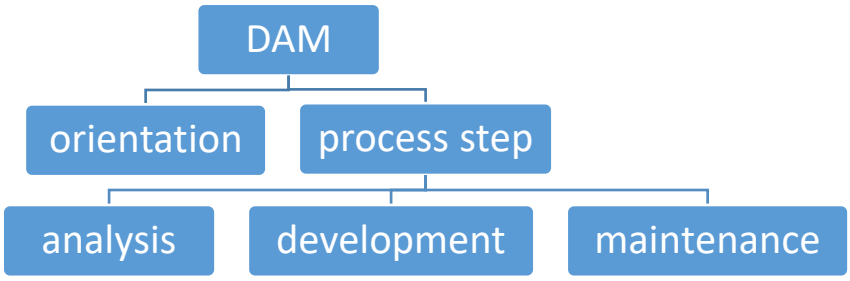


Figure 2: DAM dimension model

The DAM dimension model in Figure 2 illustrates the classes found in the literature review.

In conclusion: The literature reviews introduces many concepts which can directly be used as classes for both DAP and DAM. But there is no alignment amongst the sources, there is no common reference framework which can form a common ground. There is a great lack of a common language and understanding by what is precisely meant by DA and DAM. Synonyms found for DA are knowledge discovery data mining, knowledge discovery data analytics, data science and big data. Synonyms found for DAM include (apart from CRISP-DM): knowledge discovery analytics process model, data science project management process, team data science process, integrated knowledge discovery and data mining, refined data mining process and business analysis process. There is a clear need for harmonization and standardization. CRISP-DM has been the best reference so far but it is outdated because it does not relate to agile methodologies and it is less complete than Snail Shell. The adoption of DAMs in general is very low amongst DA practitioners, therefore the influence of CRISP-DM itself is very limited. This literature study observes all the relevant concepts and puts forward the first models for the DAP and DAM dimensions.

2.4. Objective of the follow-up research

The research model provides a structure for validation and design by follow-up research. The follow-up research intends to further clarify the DAP and DAM models and their classes. The ground works starts with understanding the space of the classes, including their subclasses. It may be helpful to use the building theory (Eisenhardt, 1989) for saturating the space of the classes. Then an important objective is to find mappings or patterns between the DAM and DAP classes.

The literature review shows that Agile practises are being increasingly applied in DAMs, for example using iterations. While the Agile methodologies are also evolving, they emphasize the need for adapting the methodology to the project by the team (self-organizing teams). If this also applies to DA, the ability to choose between different DAMs based on the project becomes more urgent. The literature study reported various DAMs but gives little guidance as to when to use a certain DAM for a certain DA project. The objective of the follow-up research is to fill this gap with a comprehensive framework. DA practitioners should be able to relate their DA project to this framework and make a rational choice for a DAM, or to continue using an ad-hoc approach. Rather than choosing a DAM, the DA practitioners may be able to compose their own DAM, linking the DAP to the DAM classes.

3. Methodology

Design Science Research (DSR) has been chosen as the methodology for this research. This chapter explains the reasons why in the sections on conceptual and technical design, followed by a reflection on rigor and relevance.

3.1. Conceptual design: select the research method(s)

DSR is an effective problem-solving based methodology for the design of artefacts to make research contributions; using evaluation, communication and scientific rigor practices (Hevner, 2007). It resolves observed problems through a design process that involves research contributions, evaluation and communication to the appropriate audience (Peffers et al., 2007).

DSR is rooted in Information Systems research. Traditional descriptive research works through exploratory, descriptive, explanatory and evaluative studies (Saunders, Lewis, & Thornhill, 2016). This is rooted in behavioural science which researches phenomena and does not necessarily work for research in Information Systems, where the objective is to innovate and design (Hevner et al., 2004). Hevner observed that it is often a stretch to find descriptive research used as a base for the creative activities of DSR. DSR produces prescriptive research in the form of research artefacts, continuously applying research rigor during the DSR cycles (Hevner, 2007).

This research often refers to DAMs, which are in itself examples of prescriptive knowledge. They prescribe methodologies and their primarily objective is to improve (and not to describe) the work of DA practitioners. This research wants to improve the design of DAMs (generating prescriptive knowledge) and therefore the DSR research paradigm is well suited.

The various research strategies are survey, case study and experiment. A survey strategy would be difficult for this research, since the DAP- and DAM conceptual space is not clear yet. The literature study demonstrated that it is beneficial to use a DAM, therefore there is no need to design an experiment to prove the value of using a DAM. A case study is most appropriate strategy here since it studies a phenomenon within a real-life context. Then there is the choice between a single or a multiple case study. The single case study is useful when a phenomenon has not yet been studied, which is not the case here. A multiple case study is useful to replicate findings across cases (Saunders et al., 2016) and has been selected for this research. The theory building concepts of Eisenhardt are useful for combining case study research with finding new concepts until so-called *theoretical saturation* has been reached (Eisenhardt, 1989).

Design Science Research (DSR) methodology has been chosen as the 'container' in which the case study will be performed. DSR helps to combine the classic case study principles of Yin (Yin, 2014) with the iterative theory development notion of (Eisenhardt, 1989). See also *paragraph 1.6 Main lines of approach* for more references on DSR. During the case study, DSR will structure the design process in order to understand the problem domain.

3.2. Technical design: elaboration of the method

This section describes the DSR steps taken in designing a new model, drawing from literature study of chapter 2. The model provides answers to the chapter 1 research questions.

Figure 3 below shows the six DSR activities (Peffers et al., 2007) mapped to the chapters of this study. Step 3 produces the artefact, which is then demonstrated, evaluated and communicated in subsequent steps and chapters. Peffers' original model shows that it is possible to iterate back to step 2 from steps 3, 5 and 6.

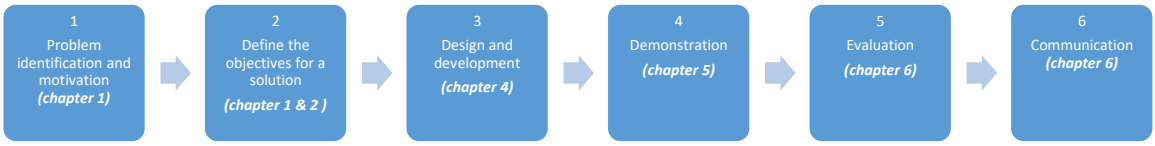


Figure 3: DSR model mapped on the chapters of this report (Peffers et al., 2007)

Step 1: DSR starts with looking at the problem that the researcher tries to resolve. Peffers et al. (2007) stresses the importance of a problem centred approach. Chapter 1 applies this with consulting literature to identify the problem and the motivation for this research.

Step 2: The objectives of the solution have been described in chapter 1, explaining also the research questions. Step 2 also provides for the theoretical framework developed in Chapter 2.

Step 3: This step further develops the theoretical framework of Chapter 2 into a design. All design choices were subject to scientific rigor and are based on the literature.

Step 4: This step demonstrates the design by deploying a multiple case study with interviews and focus groups (semi-structured following the design) with DA experts. The demonstration collects data on how organizations perform DA and how the design was perceived.

Step 5: The Framework for Evaluation in Design Science (FEDS) has been selected for the evaluation (Venable et al., 2016). The FEDS evaluation strategies are on the axes of (i) *artificial* versus *naturalistic* evaluation and (ii) *formative* versus *summative* evaluation. *Artificial* evaluation may be logical/rhetorical as the literature review in Chapter 2. *Natural* evaluation is conducted in a real-life context and this research deploys *natural* evaluation through a case study. This study applies both a formative evaluation (case study with twelve interviews followed by two focus groups) and a summative evaluation.
The *formative evaluation* uses coding techniques (open coding, axial coding and selective coding) to analyse the collected data through the process of constant comparison. The coding techniques are used to find classes, perspectives, members and groupings of members. These coding techniques are often used in the context of Grounded Theory (Saunders et al., 2016). This research uses predominantly the selective coding technique and to a lesser extend open coding and axial coding. The Ose methodology has been selected for systematic manual coding with Microsoft Excel to structure all the qualitative data (Ose, 2016).

The *summative evaluation* is detailed, using some of the criteria from the hierarchy of evaluation criteria to assess Information Systems artefacts (Prat, Comyn-Wattiau, & Akoka, 2014). The following criteria *in italics* have been chosen from the hierarchy of Prat and geared to the objectives of this study:

- *Validity* (can DA practitioners trust this artefact, is it reliable?).
- *Consistency* of the artefact with people, the organization and the technology.

- *Completeness* of the structure. Does the artefact represent the DAM and the DAP domain? Is the hierarchy and the granularity of the artefact correct?
- *Learning capability*. Can DA practitioners learn from using the artefact, does it help them with applying a DAM to their DAP?

3.3. Reflection on rigor and relevance

Scientific *relevance* in DSR is depicted as whether the (business) environment needs the research artefacts, also referred to as the 'relevance cycle' by Hevner. The environment consists of people, the organization and the technology (Hevner et al., 2004). DSR aims to produce an artefact which is relevant in a certain environment. This study articulated in chapter 1 the problem formulation and motivation to ensure its relevance. The case study has been performed with people from several organizations which were using a large variety of DA technologies to establish the right environment for the research (see chapter 5 on the DSR demonstration for details).

Scientific *rigor* in DSR is achieved through the appropriate use of the existing foundations and methodologies of what Hevner calls the 'knowledge base' (Hevner et al., 2004). The literature research (consulting the knowledge base) in chapter 2 ensures that the existing foundations are consulted properly. Chapter 2 deploys a systematic literature review, which can be repeated (Okoli & Schabram, 2010). The usage of DSR as a container for case study research is detailed here and chapter 5. There is a clear distinction in the design of chapter 4 (based only on the knowledge base) and the data which has been collected during the case study/demonstration. The iterative use of DSR as a container for case study research is detailed in in this chapter and therefore repeatable. The research model is the reference for the interviews for the case study research. The rigor is strong because of its use of well-known and well documented existing research techniques like DSR, case study research, Eisenhardt's theory building (Eisenhardt, 1989) and coding techniques from Grounded Theory (Saunders et al., 2016).

4. Design

This chapter presents the design of the DAP-DAMP matrix. This design evolved from the concepts and models presented in the literature study (Chapter 2). The design consists of models for the DAM dimension and the DAP dimension and they have a common structure. At the top level there is the so-called *dimension* (DAP or DAM), detailed with a hierarchy of classes. The structure (dimension, classes with subclasses) is important as it provides for depth in the model. This depth is missing in existing models like CRISP-DM. Whether certain classes can be grouped, whether the right hierarchy is chosen and having the choice for a certain granular level is useful for the discussion. All the classes have been chosen because they are found to be relevant in the context of choosing the right DAM. These choices are validated in the subsequent chapters, following DSR.

4.1. Matrix model

The design work in this chapter details the classes from Chapter 2 and makes mappings as shown in Figure 4. The mappings are discussed here, following the DAP classes:

Hypothesis: Projects that work on hypothesis generation have more difficulties in estimating their project completion time than projects working on hypothesis testing. The suggestion is to apply different project management processes for each of them because it can simplify project management in terms of task estimation. Finding value in data (hypothesis generation, explorative) is less straightforward for task estimation than the routine work of hypothesis testing (Saltz et al., 2017). A waterfall approach needs to have everything planned upfront. Therefore, an agile orientation is suggested for hypothesis generation projects because it is not always known whether the effort will yield results (Saltz et al., 2017). The hypothesis generation starts with the data, finding patterns in the data which then maybe useful for business understanding and/or resolving certain problems. The problem formulation is then to be interpreted as that a solution has been found for a problem of which the business was unaware it could be resolved. Hypothesis testing starts at the beginning (*problem formulation*), proceeding with the subsequent process steps. The hypothesis testing projects therefore could fit well with waterfall-oriented projects. The class orientation in the DAM model is based on work of Batra with the classes (Batra, 2018) plan-driven (in the model called *waterfall*) and agile-driven (simply called *agile* in the model).

Data: A critical success factor analysis suggests the use of iterative process models for big data projects to maximize the learning process (Gao et al., 2015). This suggestion is reflected in the linking of big data projects to an agile oriented DAM. However, big data projects have unique characteristics which require careful consideration when adopting agile methods. *Data Preparation* has technical challenges because of the volume of big data compared to the use of traditional SQL analytics. Data exploration after the data loading is not practical in a big data environment and the suggestion is to do the data exploration in its original format. During the work on data understanding, the variety, velocity, veracity and volume properties of big data are vital for identifying the challenges for modelling (Baijens & Helms, 2019). *Modelling* in big data projects tends to be done by application developers. This could result in a risk for the data-integrity because application code can be changed without consulting business users, thus affecting the data driven decision making (Li et al., 2016). The volume of big data also demands a scalable deployment environment (Li et al., 2016).

Requirements: Many publications (Ahangama & Poo, 2015; Batra, 2018; Jensen et al., 2019; Mariscal et al., 2010; Sharma et al., 2012) mention requirements by their origin (business-, customer-, project-, stakeholder-, organization requirements), but do not mention very specific classes of

requirements. To start this work, scope creep and reproducibility are identified and can initiate further research. Scope creep may lead to project failure and it is important to consider in plan-driven (waterfall) projects, together with other measures such as expectation management, contracts and risk management (Batra, 2018). Reproducibility is achieved by consistent preservation of the relevant artefacts (Saltz et al., 2017), and is relevant for the work on data preparation and modelling. The Institutional Review Board (IRB) or the Ethics Committee of the organisation and the country's data regulations will demand reproducibility of the models (Baijens & Helms, 2019; Sharma et al., 2012).

The maturity of data governance in the organization: The authors of the Snail Shell model suggest addressing enterprise data architecture in future research and increasing the maturity of the analytics capability of the organization (Li et al., 2016). Gao suggests including people, process and technology in information management strategies to overcome the challenges of big data projects (Gao et al., 2015). Mariscal uses the term data "cleansing" as part of the data preparation (Mariscal et al., 2010). Enterprise data architecture, information management strategies and data cleansing are part of the data governance and data management practices of the organization. The expectation is that the data quality is higher when an organization has an effective, highly mature data governance model in place. A project with high quality data at its disposal will proceed faster, in particular during the process steps of data understanding and data preparation (including data cleansing).

The before described linkages between the DAP and DAM classes are expressed in the below DAP-DAM matrix model with X's.

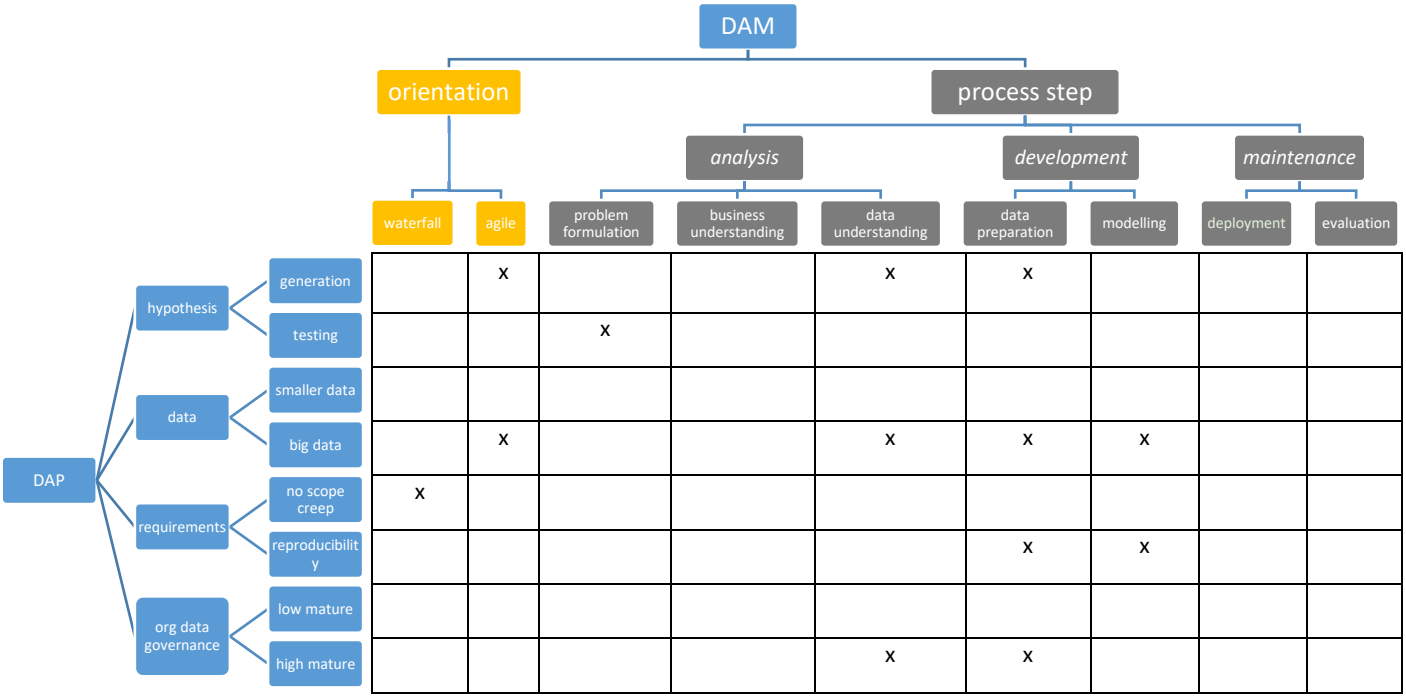


Figure 4: DAP-DAM Matrix

5. Demonstration and Evaluation

The demonstration of the artefact took place through two iterations. The first iteration consisted of twelve individual interviews. The second iteration consisted of two sessions, organized in focus groups. *The objectives of the interviews and focus groups are described separately in the following sections.* The focus groups were mainly formed from the people earlier interviewed, plus two additional experts from the UN.

UN	GIS consultant
	Geospatial data scientist
	System developer
	Information systems manager
	Information technology officer
	Data analytics and platform development
	Chief data and analytics officer

Government	Business intelligence analyst
Industrial automation	Research director
	Control systems engineer
	Innovation manager
	Application engineering manager
Research institute	Principle scientist

Table 3: Institutes & roles of the interviewees

The left column of Table 3 indicates all the roles of respondents who work for the United Nations (UN). UN personnel came from *three* different UN organizations. In total, *seven* institutes participated in *thirteen* different roles. The right column indicates the various roles per sectors. *All interviews and focus groups were held online.* The interviews and focus groups are all recorded (around 10 hours of material), transcribed and coded. Keywords were coded following the coding scheme reflected in *Appendix H Coding scheme*. Some open- and axial coding was applied, where selective coding has been applied for all the data collected. Microsoft Excel is used for coding, using the basic structure of the Ose methodology to structure qualitative data (Ose, 2016) for Microsoft Excel. Ose only works for selective coding and the structures for open- and axial coding have been added, although the selective coding was the predominantly used technique. All the data is available on request. This chapter ends with the formative and summative evaluation as methodologically laid out in chapter 3.

5.1. Iteration 1: Interviews

The *objective* of the interviews was to collect data on the current DA processes of the organizations. The interviews were all held with a videoconferencing application which gave the possibility to share the screen and to record the audio. The interviews lasted between 40-55 minutes, depending mostly on how much the respondent elaborated in their replies. The interviews were semi-structured with questions (see *Appendix G Questions* with interview questions).

The topic of whether the organization was using process models like CRISP-DM/agile/etc. was always discussed. The DAM model was only shown (time permitting) after the respondents had described their process steps. The DAP model was not shown in order not to pre-empt the discussion and because the DAP model had a less strong base in the literature than the DAM model (the DAM model is directly based on Batra, Li and Mariscal). All interviews were transcribed and analysed with merely selective coding (see Appendix H Coding scheme). To a lesser extent, open and axial coding were used to find new classes. Three interviews were held in Italian, two in Dutch and seven in English. The Italian interviews were transcribed in Italian and translated in English. The Dutch

interviews were not translated, and all coding was done in English. The data was analysed with the above mentioned Ose methodology.

5.2. Iteration 2: Focus groups

The *objective* of the focus groups was to demonstrate the designed DAP-DAM matrix and to spin-off a discussion to collect feedback data. The analyses of the data from the interviews were the base from which to iterate back to the DSR step 2, the design. A new model was designed and reflected in a presentation, which guided the focus group session. The presentation very briefly introduced DSR, the problem statement, objectives and the research questions. The Mariscal Redefined Data Mining Process Model (Mariscal et al., 2010) was shown to give the audience a feel for the dynamics and the vast scope of the conceptual domain (the Mariscal model also provided also for the first level classes of the process steps, e.g. analyses, development and maintenance). After this, the new design was introduced. The various changes were explained, and participants were continuously asked for comments and feedback. The focus group questions can be found in *Appendix G Questions*. One participant clarified his statements on augmented analytics with an email after the focus group session, and this material has also been taken into account for this research.

The focus groups were held in two sessions. The first session had a mix of all institutes with eight experts in total. The second session was conducted with just three experts from the UN. Nine experts had also participated in the interviews, and two new experts made their entrance in the first focus group. Feedback had already been collected on the DAM dimension model in the first iteration (interviews), therefore the focus groups were intended mainly to collect feedback on the DAP dimension model and on the mappings with DAM. The DAP and DAM models were shown separately as an introduction, followed by the DAP-DAM matrix. The various mappings were presented and feedback was collected on them.

5.3. Formative evaluation

The DAP classes went through four DSR iterations, see also *Appendix I: DSR iterations*. This section starts discussing why the requirements class was removed from the model. This evaluation then continues evaluating the DAP classes, making linkages with the DAM model and listing the results in tables. The new class of Self-Service-Analytics is introduced, followed by a presentation of the new DAP-DAM matrix.

Requirements: Respondents agreed that a waterfall orientation fits well with the requirement of no scope creep, and that the requirement of reproducibility links with the process step of data preparation. A general theme in industrial automation is that DA is used to fulfil the requirement of cost reduction. The requirement implementing IoT often results in finding the right digital twin for the various variables in the industrial process. The focus groups agreed that the *no scope creep* class was not useful because there are no projects without scope creep. The research of Batra also investigated scope creep and other aspects in relation to project success and concluded that there was no significant effect (Batra, 2018). The class reproducibility was found to be relevant in general, but not in relation to the DAM. The concept of reproducibility is part of the data governance, which is already in the model. Therefore, the DAP class *requirements* was removed from the model.

DAP class	DAM class	Status	Comments
Requirements- no scope creep	Waterfall	Rejected	There is always scope creep.
Requirements- reproducability	Data preparation	Rejected	Covered by data governance.
	Modelling	Rejected	

Table 4: Requirements

Hypothesis generation: DA in industrial automation often starts with hypothesis generation in a simulated way. Simulated means here that certain IoT sensors with associated microservices are simulated though a high-level cloud function. When the hypothesis is proven to be valuable, the sensor is acquired, and the high-level cloud function is implemented as a microservice. Another evolution of hypothesis generation into hypothesis testing is where research institutes are asked to work on hypothesis generation. When successful, the hypothesis is operationalized in the field through hypothesis testing. The respondents confirmed that agile is more suitable for hypothesis generation because it allows for coping with uncertainties. They also confirmed that hypothesis generation starts with Data Understanding and that hypothesis testing starts with Problem Formulation. The focus group pointed out that hypothesis generation also needs to be linked with business understanding.

Hypothesis generation versus orientation (waterfall/agile). The initial tendency of applying agile to hypothesis generation was detailed in the focus groups. The choice for agile depends on other factors like whether the project could carry the cost of a full-fledged agile implementation including continuous integration. As with the interviews, there were mixed interpretations of waterfall and agile. The term agile alone creates confusion. Some use it to refer to agile as an iterative process, while others associate it with its formal adoption through the use of SCRUM or Kanban. Others think of agile in terms of its practices like continuous integration and it depends on the agile preparedness of the organization whether this can be chosen for a project. Using agile practices in DAPs is not new in the literature (Baijens & Helms, 2019). Therefore, an extra DAM class has been defined under orientation with the name *iterative* to make the distinction from a formal agile approach more explicit. Various respondents spoke about the concept of a *pipeline,* often in the context of discussing iterative/agile practices. The pipeline can be understood as the programmatical declaration of the process steps ranging from data preparation to deployment. The pipeline is monitored through a continuous integration process, where the data and models are frequently tested and deployed.

Hypothesis testing: Descriptive analytics projects were mainly working on hypothesis testing. Hypothesis testing is used in the EU and UN for policy monitoring and disaster impact analysis through earth observation with satellite imagery.

DAP class	DAM class	Status	Comments
Hypothesis generation	Iterative	New	Making the difference explicit with formal agile methodology adoption.
	Agile	Accepted	Associated with continuous integration and with formal adoption of an agile methodology.
	Business understanding	Added	The work starts with business understanding, else the exercise doesn't make sense.
	Data understanding	Rejected	Confirmed as the next step, rejected as the first step.
Hypothesis testing	Problem formulation	Accepted	The work starts with problem formulation.

Table 5: Hypothesis

Big data: The respondents confirmed that big data projects are more efficient with an agile orientation. The respondents agreed with the class in general. One respondent indicated that talking about big data in terms of volume would be too narrow and that the number of data sources also needed to be considered. This confirms the need to apply the four V's (variety, velocity, veracity and volume) as explained in chapter 2. One respondent mentioned the term *big data cloud-native* in

relation to the democratisation of data. This refers to the phenomena that the cloud is bringing DA to a higher level (thus becoming more accessible for a wider audience): the cloud helps with the heavy lifting during the process steps of data understanding, data preparation, modelling and deployment. The cloud provides for flexibility and decreases the workload. Non tech-savvy DA practitioners are thus empowered to perform DA without being blocked by big data infrastructural concerns. The cloud stimulates more organisations to work with big data and decreases the barriers to working with big data. Cloud-native applications can run only in the cloud. Cloud architecture patterns are fast emerging, from Infrastructure-As-A-Service (IAAS) to Function-As-A-Service (FAAS). This confirms the trend that the cloud helps with the heavy lifting, moving DA away from the physical level to the functional level. On the downside, cloud-native increases a vendor lock-in and it can result in significantly higher costs (Kratzke & Siegfried, 2020).

Smaller data: The smaller data DA is active in the field of descriptive analytics and increasingly uses self-service architectures. Self-service architectures lift the burden of data understanding and data preparation because the data has already been loaded, standardized and made available. One respondent in the industrial automation sector mentioned the cloud architecture pattern of Plant-as-a-Service in relation to self-service. This implies that the machines and sensors are managed and serviced through a self-service model, following cloud-native patterns. Smaller data is therefore moving in a similar direction to big data as discussed beforehand.

DAP class	**DAM class**	**Status**	**Comments**
Smaller data	Problem formulation	Accepted	Smaller data projects start the work with problem formulation. Smaller data tends to occur more in hypothesis testing/descriptive analytics projects, regardless its orientation.
Big data	Agile	Accepted	Confirmed as more appropriate for big data projects.
	Iterative	New	
	Data understanding	Added	Big data projects require extra attention for these three process steps. Practices here are increasingly subject to cloud (native) trends, making big data DA is a fast-changing discipline.
	Data preparation	Accepted	
	Modelling	Added	

Table 6: Smaller data – big data

Organisational data governance: The demonstration confirmed the design that a high mature data governance level in an organisation positively affects the work on data understanding and data preparation. One respondent mentioned that the existence of a data lake is a sign of DA readiness of the organization. Another respondent reported that imposed data governance practices within an institute would only lead to less data. Apart from the above mentioned respondent, there was consensus that the maturity of information management significantly influences the project. It was pointed out that data governance is particularly important for self-service analytics. It creates the necessary trust when the data owners are well defined. Small and medium sized enterprises generally tend to have a low level of data governance maturity. Various interviewees reported on the difficulty of doing proper information management. It was often not clear whether information management was associated with the amount of available data within the organization, the availability of tools to manage the data or to the level to which the data was harmonized/standardised. This confirms that data governance is associated with many aspects. The omnipresence of data, its volume and its intrinsic value resulted in a need to manage the data explicitly, as formulated by (Brous, Janssen, & Vilminko-Heikkinen, 2016): "*More and more data is*

becoming available and is being combined which results in a need for data governance - the exercise of authority, control, and shared decision making over the management of data assets".

DAP Class	DAM class	Status	Comments
Org data governance - Low mature	All	Accepted	All steps potentially take longer, suffering from low data quality.
Org data governance - High mature	Data understanding	Accepted	These process steps benefit in particular a lot from comprehensive high-quality data as a result from proper organizational data governance.
	Date preparation	Accepted	
	Modelling	Added	

Table 7: Data governance

New DAP class (SSA): The demonstration resulted in the discovery of a new class: *Self-Service Analytics (SSA).* The class came out in direct and indirect formulations. A dashboard is often the first step in the evolution as a project to a self-service model. Descriptive analytics projects often used the term SSA explicitly. One institute referred to Palantir (a private American DA software company) which offers a maturity framework with five levels of self-service maturity. SSA was explained as providing an DA architecture with at least two layers, the base layer and the self-service layer. The base layer is developed by a separate team and consists of data and tools. The team consists of DA experts with solid data and IT skills. Data is preloaded, aggregated and harmonized. This base layer is then offered to the business in the organization which can use this in a self-service mode. Another organisation had a similar breakdown, but only for the data. Data is prepared at HQ-level and then used in the field in various geographical locations by DA experts. The literature confirms that SSA is an important trend. SSA allows business users to perform DA while being less dependent on DA experts and is therefore becoming a top priority for organizations (Daradkeh, 2019). SSA greatly speeds up the process of *modelling* and democratizes analytics (Schuff, Corral, St. Louis, & Schymik, 2018). Big data DA uses cloud native tools where the data is already preloaded and as such is also moving to a self-service approach. Recent research in healthcare refers to self-service data science following an evolutionary model from foundational-, applied-, self-service- to citizen data science. Self-service data science is only applicable for the CRISP-DM *data understanding* process step (Ooms & Spruit, 2020). This observation however applies to the field of automated machine learning (predictive analytics) and may not be generalizable to descriptive and prescriptive analytics. SSA is linked in the DAP-DAM matrix to the process steps *data understanding* and *modelling*.

DAP class	DAM class	Status	Comments
With SSA	Data understanding	Accepted	Offering SSA to the organization requires preparing a base layer of data which is easily understood.
Without SSA	Modelling	Accepted	Modelling is labour intensive without SSA.

Table 8: SSA

New DAM class Iterative: As discussed earlier, the agile class is too coarse grained and is completed with the iterative class. Regarding the maintenance class, some respondents initially had difficulties with its name, not with its conceptual position nor with the sub classes of deployment and evaluation. However, no better name was proposed during the discussions.

The evaluation resulted in a new matrix as shown in Figure 5 (next page). The x's are the existing linkages as discussed in the DSR design step. The n's are the new linkages evaluated in this section.

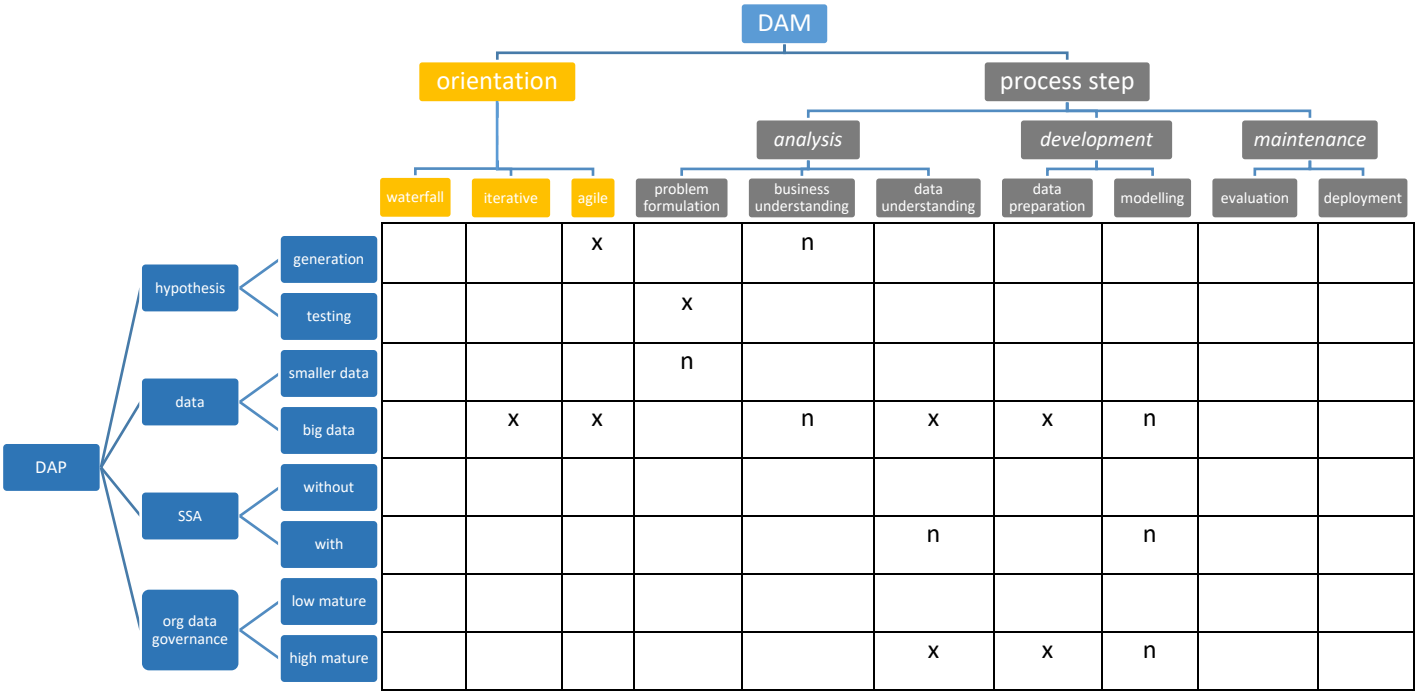


Figure 5: DAP-DAM Matrix Revised

5.4. Summative evaluation

Section 3.2 discusses the method of the summative evaluation based on the hierarchy of criteria for information system artefact evaluation (Prat et al., 2014). Four criteria were selected, and the subsequent evaluation is performed here:

Goal validity: The degree to which the artefact works correctly. The initial model was based on literature research and DSR design, always profoundly grounded in what DSR calls the scientific knowledge base. All design decisions are corrected and completed during the DSR iterations. Following the DSR methodology, four iterations were performed, as depicted in *Appendix I: DSR iterations*.

Consistency with the environment (people, organization and technology): 'Environment' here is defined according to definitions used for the environment of information systems (Hevner et al., 2004). In terms of people, a total of 14 DA practitioners were consulted. The below table illustrates the diversity of the environment:

people	2 senior managers versus 12 practitioners
	7 descriptive analytics versus 7 both (descriptive/predictive analytics)
	13 different roles*
organization	7 institutes*
	4 sectors*
technology	6 Geospatial Analytics
	2 Data warehousing
	2 Self-Service-Analytics
	4 IoT (industrial automation)

** See table 3 in chapter 4*

Table 9: Consistency with environment

The intent of conducting a multiple case study is achieved successfully given the number of seven institutes. The distribution of people between management and practitioners is unbalanced. This

may explain why the DAM evolved less during the DSR iterations because there were only two managers, and managers are responsible for the process. The technology distribution is good, the geospatial analytics practitioners are however overrepresented, but this did not dominate any discussion. A possible missing DAP class is the notion of descriptive and predictive analytics (and possibly prescriptive analytics).

Completeness: The DAP dimension of the model is complete, it was extensively (re-)designed and (re-)discussed, and the same was true for the orientation class of the DAM dimension. The process steps of the DAM were judged to be complete and were therefore not subject to discussions with the exception of the maintenance class. It is likely that the maintenance class, along with the subclasses of evaluation and deployment within the DAM dimension model, is not complete/correct, and there are missing linkages with the DAP dimension.

Learning capability: The model reflects the conceptual space of DAPs and DAMs effectively, including its linkages, and this is very educational. It provides practitioners the vocabulary to reflect on their actual projects at hand. The centre of gravity of this research is on the DAP dimension model and the DAM class orientation. The model is less articulated with regard to the process steps, but that part is heavily rooted in existing models to which the interested reader is referred.

6. Discussion, recommendations, conclusions and reflection

The results of this research are reflected in the evaluation of the previous chapter, and ultimately in the presentation of the revised model. This chapter extends on this with discussions, recommendations, conclusions and reflections.

6.1. Discussion and recommendations

The literature study in Chapter 2 clarified that the centre of gravity of DAM literature is focussed on the DAMs themselves and less on the various DAP classes and its linkages. Instead, this research contributes to an overarching perspective through the two dimensions, DAP and DAM. The current model allows for visual clear mapping DAP to DAM classes.

The developed model has currently no connections between the DAP and DAM classes on the maintenance class (with evaluation and deployment). A couple of respondents talked about the concepts of pipeline and continuous integration in the context of maintenance, evaluation and deployment. Continuous integration intends to manage the DA pipeline, including evaluation (continuous testing) and (continuous) deployment. Some refer to this as agile practices or a DevOps mind- and toolset (Steinwandter, Borchert, & Herwig, 2019). The *Appendix J: Pipeline and Continuous Integration* shows that these concepts are mentioned more significantly in combination with predictive analytics. This may indicate that there is a need for a new DAP class which distinguishes between descriptive- and predictive analytics. Descriptive analytics is more associated with traditional data warehouses and predictive analytics more with data science. The recommendation is to investigate whether distinguishing descriptive from predictive analytics produces significant insights that enrich the current model. The impact of the difference between descriptive and predictive analytics on the DAM is not clear enough. There are signs that the practices are converging but in reality are still very distinct (apples and pears?), not to mention prescriptive analytics.

The DAP and the DAM dimension models themselves resemble the ontology BIGOWL (Barba-González et al., 2019). BIGOWL is an ontology to support knowledge management in Big Data analytics. BIGOWL defines three ontologies, (i) use case ontology, (ii) algorithm ontology and (iii) workflow ontology. These three ontologies bear a resemblance to the three classes of Mariscal used in the DAM dimension model (analyses, development and maintenance). The model in this research shares with an ontology the idea of depth through classes and subclasses. The three BIGOWL ontologies are essentially perspectives on DA. CRISP-DM and Snail Shell give only a process step perspective on DA. This research gives two layered perspectives, DAP and DAM. The three BIGOWL perspectives may reveal valuable new insights. The recommendation is to investigate whether an ontology approach would help in understanding the semantics of the DAP and DAM dimensions, including its linkages.

The DAM dimension model relies heavily on the Snail Shell model (Li et al., 2016). The Snail Shell model has only one class more (problem formulation) than CRISP-DM. The class of problem formulation however is fundamental in illustrating the difference between the starting points of hypothesis generation and hypothesis testing projects.

The demonstrations led to the discovery of the class Self-Service Analytics (SSA). It is becoming a top priority for organizations (Daradkeh, 2019). SSA greatly speeds up the process of modelling and democratizes analytics (Schuff et al., 2018). SSA started off mainly in the field of descriptive analytics but is increasingly applied in the field of predictive analytics (Ooms & Spruit, 2020), however its conceptual meaning differs per field. Two recent sources refer to the phenomena of SSA by the

name of Augmented Analytics, describing it as artificial intelligence that helps the DA-practitioners in their work of going through the DAM process steps (Andriole, 2019; Prat, 2019). Research on DAMs need to distinguish the fields of descriptive- and predictive analytics *or* harmonize terminology over the two fields. In addition, SSA and Augmented Analytics are excellent candidates as new DAP classes to take on board in order to ensure that research keeps pace with the fast-changing DA discipline.

6.2. Conclusions

The main research question (What instruments are needed to convince DA practitioners to use a DAM?) is implicitly answered by providing the DAM-DAP-matrix as a tool for making informed decisions. It helps us to understand how the DAP reality can be mapped to process steps and methodologies. The sub research questions on classes and mappings are answered by extensive research into the relevant classes and mappings. The DSR steps of demonstration and evaluation resulted in a new DAP-DAM-matrix. Four DAP top level classes are identified (hypothesis, data, self-service analytics and data governance). Two DAM top level classes are identified (orientation and process steps). Fourteen DAM-DAP mappings are identified.

The literature study resulted in the identification of hypothesis generation/testing, big data/smaller data and data governance as main classes for a DAP. Following DSR methodology, the self-service analytics was discovered as the missing DAP class during the DSR demonstration and evaluation steps. The literature study revealed waterfall/agile as an important DAM orientation classes. The demonstration revealed that there is confusion on the meaning on what can be called an agile project. The missing class *iterative* was added to the list of orientation to clarify that all agile methodologies are iterative but not all iterative methodologies are agile.

The limitation of the designed DAP-DAM matrix is in the last process step of maintenance. No DAP mappings are articulated for that step and its composition needs more research. The maintenance step has two fine-grained steps of evaluation and deployment. The demonstration revealed doubts on this composition, but no improvements came through.

The theoretical value of this research two-fold: (i) the contextualization, design and articulation of DAM classes with DAP classes (including their mappings), (ii) the acknowledgement that DAM classes are not just a series of process steps but they are subject to wider methodological orientations such as waterfall, iterative and agile. This will help the DA practitioners in practical terms to compose a specific DAM for the DAP they have at hand.

6.3. Reflection

The DSR methodology was found to be valuable because of its capability to produce new prescriptive knowledge (the design). This research went through three DSR iterations (interviews and two focus groups). There was not enough time to fully analyse and design a new artefact before the next iteration. As a result, the new class of self-service analytics was only demonstrated in its current form to the second focus group. This case study did not reach theoretical saturation as envisaged (Eisenhardt, 1989), but it managed to perform the iteration of twelve interviews followed by two focus group sessions.

The introduction of this research explains the three main categories of DA, descriptive-, predictive- and prescriptive analytics (Sharda et al., 2018). This research did efforts to reflect their differences through the classes smaller/big data and hypothesis generation/testing. Some respondents did comment that the DAM was merely reflected predictive analytics and/or data science. The limitation

of this research is the upfront generalization of the three categories into DA only and this should be validated in future research.

A challenge during the demonstration was to have the respondents focussing reflecting on classes and mappings without pre-empting the discussion and/or going into details. Since the adoption level and maturity level of methodologies in DA is low, also the practitioners had difficulties in articulating the classes and mappings. Only after a while the researcher was more able to gear the discussion in the right direction in order to extract the needed data.

References

Abdalhamid, S., & Mishra, A. (2017). Adopting of agile methods in software development organizations: Systematic mapping. *TEM Journal*. https://doi.org/10.18421/TEM64-22

Ahangama, S., & Poo, D. C. C. (2015). What methodological attributes are essential for novice users to analytics? – an empirical study. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-319-20618-9_8

Andriole, S. J. (2019). Artificial Intelligence, Machine Learning, and Augmented Analytics. *IT Professional*. https://doi.org/10.1109/MITP.2019.2941668

Baijens, J., & Helms, R. W. (2019). Developments in knowledge discovery processes and methodologies: Anything new? *25th Americas Conference on Information Systems, AMCIS 2019*.

Barba-González, C., García-Nieto, J., Roldán-García, M. del M., Navas-Delgado, I., Nebro, A. J., & Aldana-Montes, J. F. (2019). BIGOWL: Knowledge centered Big Data analytics. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2018.08.026

Batra, D. (2018). Agile values or plan-driven aspects: Which factor contributes more toward the success of data warehousing, business intelligence, and analytics project development? *Journal of Systems and Software*. https://doi.org/10.1016/j.jss.2018.09.081

Brous, P., Janssen, M., & Vilminko-Heikkinen, R. (2016). Activities: A Systematic Review of Data Governance Principles. *IFIP International Federation for Information Processing 2016*. https://doi.org/10.1007/3-540-68339-9_34

Colas, M., Finck, I., Buvat, J., Nambiar, R., & Raj Singh, R. (2014). Cracking the Data Conundrum : How Successful Companies Make Big Data Operational. *Capgemini Consulting*, 17. Retrieved from https://www.capgemini.com/consulting/wp-content/uploads/sites/30/2017/07/big_data_pov_03-02-15.pdf%0Ahttps://www.capgemini-consulting.com/resource-file-access/resource/pdf/cracking_the_data_conundrum-big_data_pov_13-1-15_v2.pdf

Daradkeh, M. (2019). Determinants of self-service analytics adoption intention: The effect of task-technology fit, compatibility, and user empowerment. *Journal of Organizational and End User Computing*. https://doi.org/10.4018/JOEUC.2019100102

Eisenhardt, K. M. (1989). Building Theories from Case Study Research. *The Academy of Management Review*. https://doi.org/10.2307/258557

Gao, J., Koronios, A., & Selle, S. (2015). Towards a process view on critical success factors in Big Data analytics projects. *2015 Americas Conference on Information Systems, AMCIS 2015*.

Hevner, A. R. (2007). A Three Cycle View of Design Science Research. *Scandinavian Journal of Information Systems*. https://doi.org/http://aisel.aisnet.org/sjis/vol19/iss2/4

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly: Management Information Systems*. https://doi.org/10.2307/25148625

Jensen, M. H., Nielsen, P. A., & Persson, J. S. (2019). Managing Big Data Analytics Projects: The Challenges of Realizing Value. *Proceedings of the 27th European Conference on Information Systems (ECIS)*.

Kisielnicki, J., & Misiak, A. M. (2016). Effectiveness of agile implementation methods in business intelligence projects from an end-user perspective. *Informing Science*.

https://doi.org/10.28945/3515

Kratzke, N., & Siegfried, R. (2020). Towards cloud-native simulations – lessons learned from the front-line of cloud computing. *Journal of Defense Modeling and Simulation*. https://doi.org/10.1177/1548512919895327

Lepenioti, K., Bousdekis, A., Apostolou, D., & Mentzas, G. (2020). Prescriptive analytics: Literature review and research challenges. *International Journal of Information Management*, *50*, 57–70.

Li, Y., Thomas, M. A., & Osei-Bryson, K. M. (2016). A snail shell process model for knowledge discovery via data analytics. *Decision Support Systems*. https://doi.org/10.1016/j.dss.2016.07.003

Mariscal, G., Marbán, Ó., & Fernández, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *Knowledge Engineering Review*. https://doi.org/10.1017/S0269888910000032

Mousannif, H., Sabah, H., Douiji, Y., & Oulad Sayad, Y. (2016). Big data projects: just jump right in! *International Journal of Pervasive Computing and Communications*. https://doi.org/10.1108/IJPCC-04-2016-0023

Okoli, C., & Schabram, K. (2010). Working Papers on Information Systems A Guide to Conducting a Systematic Literature Review of Information Systems Research. *Working Papers on Information Systems*. https://doi.org/10.2139/ssrn.1954824

Ooms, R., & Spruit, M. (2020). Self-Service Data Science in Healthcare with Automated Machine Learning. *Applied Sciences*, *10*(9), 2992.

Ose, S. O. (2016). Using Excel and Word to Structure Qualitative Data. *Journal of Applied Social Science*. https://doi.org/10.1177/1936724416664948

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*. https://doi.org/10.2753/MIS0742-1222240302

Prat, N. (2019). Augmented Analytics. *Business and Information Systems Engineering*. https://doi.org/10.1007/s12599-019-00589-0

Prat, N., Comyn-Wattiau, I., & Akoka, J. (2014). Artifact evaluation in information systems design-science research - A holistic view. *Proceedings - Pacific Asia Conference on Information Systems, PACIS 2014*.

Saltz, J., Hotz, N., Wild, D., & Stirling, K. (2018). Exploring project management methodologies used within data science teams. *Americas Conference on Information Systems 2018: Digital Disruption, AMCIS 2018*.

Saltz, J., Shamshurin, I., & Connors, C. (2017). Predicting data science sociotechnical execution challenges by categorizing data science projects. *Journal of the Association for Information Science and Technology*. https://doi.org/10.1002/asi.23873

Saunders, M., Lewis, P., & Thornhill, A. (2016). *Research Methods for Business students.(ed. 7 th) Harlow*. Pearson Education Limited.

Schuff, D., Corral, K., St. Louis, R. D., & Schymik, G. (2018). Enabling self-service BI: A methodology and a case study for a model management warehouse. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-016-9722-2

Sharda, R., Delen, D., Turban, E., Aronson, J. E., Liang, T.-P., King, D., … Kong, H. (2018). *Business*

Intelligence, Analytics, and Data Science: A Managerial Perspective. Retrieved from www.pearsonglobaleditions.com

Sharma, S., Osei-Bryson, K. M., & Kasper, G. M. (2012). Evaluation of an integrated knowledge discovery and data mining process model. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2012.02.044

Spoelstra, J., Zhang, H., & Kumar, G. (2016). *Data Science Doesn't Just Happen, It Takes a process*. Retrieved from https://channel9.msdn.com/Events/Machine-Learning-and-Data-Sciences-Conference/Data-Science-Summit-2016/MSDSS23

Steinwandter, V., Borchert, D., & Herwig, C. (2019). Data science tools and applications on the way to Pharma 4.0. *Drug Discovery Today*. https://doi.org/10.1016/j.drudis.2019.06.005

Tria, F. Di, Lefons, E., & Tangorra, F. (2018). A Framework for Evaluating Design Methodologies for Big Data Warehouses: Measurement of the Design Process. *International Journal of Data Warehousing and Mining*. https://doi.org/10.4018/IJDWM.2018010102

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: a framework for evaluation in design science research. *European Journal of Information Systems*, *25*(1), 77–89.

Wixom, B. H., Yen, B., & Relich, M. (2013). Maximizing value from business analytics. *MIS Quarterly Executive*.

Yin, R. K. (2014). Case study research: Design and methods (5th ed.). In *Thousand Oaks, CA: SAGE Publications*.

Appendix A Base articles

This is the list of articles provided by the tutor during the Master thesis course of the Open University of the Netherlands in in September 2019:

- (Ahangama & Poo, 2015)
- (Baijens & Helms, 2019)
- (Gao et al., 2015)
- (Jensen et al., 2019)
- (Li et al., 2016)
- (Mariscal et al., 2010)
- (Saltz et al., 2017)
- (Saltz et al., 2018)

The tutor (Jeroen Baijens MSc) and the second reader (prof.dr.ir. Remko Helms) of this master thesis are the authors of the second mentioned article.

Appendix B Query development

The query development started with reporting on the number of results found while searching on the topic. Later on, the query was only used for searching on title. The query is developed using Web of Science, as mentioned in Chapter 2, observing only articles from 2010-2019.

Query	**Number of results**
(bigdata OR “big data” OR data) (analytics OR science) project (process OR methodology)	2488
(bigdata OR “big data” OR data) (analytics OR science) project methodology	785
(bigdata OR “big data” OR “data analytics” OR “data science”) project methodology	163
(bigdata OR “big data” OR “data analytics” OR “data science”) project (process OR methodology)	586
(bigdata OR “big data” OR “data analytics” OR “data science”) project (process OR methodology OR method)	797
(bigdata OR “big data” OR “data analytics” OR “data science”) project (process OR methodology OR method) agile	13
(bigdata OR "big data" OR "data analytics" OR "data science") ("project process" OR "project methodology" OR "project method")	1
("data mining" OR analytics OR bigdata OR "big data" OR "data analytics" OR "data science") ("project process" OR "project methodology" OR "project method" OR "process view" OR "methodological" OR "methodologies" OR "process models")	1760

("data mining" OR analytics OR bigdata OR "big data" OR "data analytics" OR "data science") ("project process" OR "project methodology" OR "project method" OR "process view" OR "methodologies" OR "process models")	1009
("data mining" OR analytics OR bigdata OR "big data" OR "data analytics" OR "data science") ("project process" OR "project methodology" OR "project method" OR "process view" OR "methodology" OR "process model")	3810
("data mining" OR analytics OR bigdata OR "big data" OR "data analytics" OR "data science") ("project process" OR "project methodology" OR "project method" OR "process view" OR "methodology" OR "methodologies" OR "process model")	4554 3715 (last 5 years) 83 (title)
("data mining" OR analytics OR bigdata OR "big data" OR "data analytics" OR "data science") ("project process" OR "project methodology" OR "project method" OR "process view" OR "methodology" OR "methodologies" OR "methodological" OR "process model" OR "project" OR "projects")	6726 (topic) 219 (title)

Query Discrepancy: The query ends with the search terms project and projects. These terms are already used in this part of the query: *"project process" OR "project methodology" OR "project method"*. Therefore, they are expected not be necessary to use separately. When they are removed in the combination terms like this:

("data mining" OR analytics OR bigdata OR "big data" OR "data analytics" OR "data science") ("process" OR "methodology" OR "method" OR "view" OR "methodologies" OR "methodological" OR "process model" OR "project" OR "projects")

It results in 872 records which is not workable from a practical point of view. When they are removed in as separate terms like this:

("data mining" OR analytics OR bigdata OR "big data" OR "data analytics" OR "data science") ("project process" OR "project methodology" OR "project method" OR "process view" OR "methodology" OR "methodologies" OR "methodological" OR "process model")

This results in only 149 documents, which is too small as an input for the literature review.

When removing project but leaving projects:

("data mining" OR analytics OR bigdata OR "big data" OR "data analytics" OR "data science") ("project process" OR "project methodology" OR "project method" OR "process view" OR "methodology" OR "methodologies" OR "methodological" OR "process model" OR "projects")

This results in 176 documents which is also too small as an input for the literature review. Therefore, the choice is to continue with this apparent discrepancy and to continue with the original developed query.

Appendix C Quality Appraisal

	Qualitative	Quantitative	Quality of deployed research method
(Ahangama & Poo, 2015)		X	Well-structured research. Many hypothesis (7) which make the research less focused and complicates the research model.
(Baijens & Helms, 2019)	X		Excellent research with useful overview of process model step. The research is an extended literature study.
(Batra, 2018)		X	Well-structured research. Many hypothesis (9) which make the research less focused and complicates the research model. Very well described methodology.
(Gao et al., 2015)	X	X	Useful research with fairly clear steps. Does not take existing DAM research into account. Very much focused on People, Process and Technology which limits the generic value.
(Jensen et al., 2019)	X		Single case study research with the useful perspective of Benefits Realization Management. It is weak on generalization because only once case is researched.
(Li et al., 2016)	X		Good research with clear steps. Strong emphasize on the presented snail shell model, which decreases its generic contribution.
(Mariscal et al., 2010)	X		Excellent overview of methodologies. New model introduced is less relevant and extensive discussion is missing.
(Saltz et al., 2017)	X		Excellent research with clear steps like literature review, research, discussion and conclusion.
(Saltz et al., 2018)		X	Short but excellent research with clear and useful outcomes.
(Sharma et al., 2012)	X	X	Good research with clear steps. Strong emphasise on the presented IKDDM model, which decreases its generic contribution.

Appendix D Data Extraction

Note: Also the article 'Effectiveness of Agile implementation methods business intelligence projects' (Kisielnicki & Misiak, 2016) was analysed because of its promising title. This article was not part of the practical screen results and was found via the Mendeley automatic suggestions. The article is however not relevant because it concludes only that Agile works well for BI.

Author	*Referred DAMs*	*Referred Methodologies*	*DAM concepts*	*DA Project concepts*	*Other concepts*
(Ahangama & Poo, 2015)	•CRISP-DM •SEMMA	•Technology Acceptance Model (TAM) •Theory of Diffusion of Innovation (DoI)	•**Relative Advantage** •Compatibility •**Result Demonstrability** •Triability •Usefulness •Knowledge Management (KM)	•Project Management (PM) •PM usefulness	Model Characteristics: - ease of use - relative advantage – compatibility - result demonstrability – try-ability Supporting Elements: - PM usefulness - KM usefulness Research model: Model Characteristics and Supporting Elements affect Intention to use Significant factors: Relative advantage and Result demonstrability
(Baijens & Helms, 2019)	•CRISP-DM •Mariscal et al. (2010)	Agile: Agile practices: •Continuous integration •Pair programming •Sprint efforts •Stand up meetings •Test driven development •Time-boxed iterations •User story SCRUM Kanban	Steps Tasks Agile practices	Stakeholders	Mapping of 5 models to the 17 steps of Mariscal. Then identifies tasks per process step of Mariscal. Then every CRISP-DM process step is mapped to Agile practices. Developments in DAMs are described along the lines of three directions. tasks, steps and agile practices Tasks are a grouped per step. The research maps agile to CRISP-DM, not to the 17 steps of Mariscal.
(Batra, 2018)		•RUP •SCRUM Agile manifesto	•agile values •plan-driven •agile-plan balanced •agile heavy	•technological capability •shared understanding •top management commitment •complexity •project success •development method •project duration •size •development method •organizational culture.	agile-heavy agile-plan -technological capability -shared understanding - top management commitment - complexity - agile values - plan driven aspects - project success DW/BIA

(Gao et al., 2015)		•Big Data Strategy	•6 Phases: business, measurement, data, learning, implementation, analysis	•6 Critical Success Factors (CSF) 1 Identifiable business value 2 Innovative analysis tools 3 Adequate hardware 4 Analytical skillset 5 Information strategy for big data 6 Big data as strategic instrument •PPT: People, Process, Technology	- 6 out of 27 success factors were declared mission critical - 55% of big data projects don't get completed - People, Process and Technology RQ How can organizations embrace success in Big Data text analytical projects? sub RQ - which process model can be applied to Big Data projects? - What are critical success factors for Big Data projects? - Which role do individuals critical success factors play at different project stages? CSF A level: - investment in the needed and novel tools - flexible IT infrastructure - scalability - focus on the business value of the projects - availability of analytical talents within the organizations - working in multidisciplinary teams CSF C level - outsourcing
(Jensen et al., 2019)	CRISP-DM	Benefits Realization Management (BRM)	Process of extracting insights from BDA can be broken down into five stages: (1) acquisition and recording (2) extraction, cleaning, and annotation (3) integration, aggregation, and representation; (4) modelling and analysis (5) interpretation (Labrinidis and Jagadish, 2012) Recommened research questions: -Project methodology support (RBM including monitoring) -Balance costs of comprehension and ease-of-use -Assessing users' understanding	1 Formulate overall business case and prioritize 2 Appreciate organizational context 4 Explicate overall benefits 4 Define benefits measures 5 Understand benefits relationships across departments 6 Measure of benefits and usefulness for end-users 7 Manage missing benefits 8 Establish end-users	
(Li et al., 2016)	•CRISP-DM •SEMMA •KDDM •KDDA •Agile Analytics •Snail Shell model	•Value Focused Thinking (VFT) •GoalQuestionMetric (GQM) •SMART •Analytics Capability Maturity (ACM)	Snail shell model 8 key phases: 1 problem formulation 2 business understanding 3 data understanding 4 data preparation 5 modelling 6 evaluation 7 deployment 8 maintenance Implements: 1) the need for problem formulation phase for establishing realistic expectations of analytic outcomes	No project elements	Techniques as Value Focused Thinking (VFT), Goal Question Metric (GQM) and SMART = specific, measurable, achievable, relevant and time-bounded. Decision style maturity: dynamic decision style model: 1) information use 2) focus Decision patterns: 1) satisfier 2) maximizer. Focus: 1) unifocus 2) multifocus Decision style model combines the four dimensions.

			2) the need for model management phase to monitor, update and/or retire models in a timely manner 3) flexibility to move between phases during the KDDA process.		The paper follows Gregor and Hevner design research study publication schedule. The research only had three high level analytics maturity areas: organization, data and decision style. Gartner has 5: unaware, opportunistic, standards, enterprise and transformative. Agile Analytics is gaining significant popularity: A value-driven approach to BI and DWH.
(Mariscal et al., 2010)	•KDD •CRISP-DM •SEMMA •Cabena •Two crows •Anand and Buchner •Cios •Marban •Six-sigma •DMIE •Marban	Six Sigma	•Process model •Paradigm •Methodology •Lifecycle •Process versus Methodology RDMP: 1 analysis 2 development 3 maintenance 17 subprocesses extracted from analyzed methodologies and process models: life cycle selection, domain knowledge elicitation, human resource identification, problem specification, data prospecting, data cleaning, preprocessing, data reduction and projection, choosing the data mining function, choosing the data mining algorithm, Build model, Improve model, evaluation, interpretation, deployment, automate, and establish on-going support.		•Knowledge Discovery Databases (KDD) Explains clear difference between process model and methodology. Process models define what to do, methodologies define how to do. Uses the term encoding schemes for classification. Discusses a human centred approach. Suggesting that many of the processes in software engineering are important for developing any type or DM engineering model. Refers to McCall to explain a good process model: effective, maintainable, predictable, repeatable, quality, traceable. The new model presented is: Refined Data Mining Process
(Saltz et al., 2017)	•CRISP-DM		•Data Science Project Model: Infrastructure, Discovery •Data Science Project Management Process	•Data, Analytical, Team and Organizational context •4Vs: Volume, Variety, Velocity, Veracity •Size of Org •**Hypothesis** generation versus hypothesis testing •**Infrastructure** (Vs) versus **Discovery** (type of analysis vs culture)	This article provides many links between Scrum, Kanban CRIPS-DM for the Data Science domain. Saltz (2017): collection, preparation, visualization, management and preservation. Having no process in place: - thwarting efficiency / improvement team - slow information sharing - delivering the wrong thing - lack of reproducibility - poor coordination - scope creep Scrum, Kanban and CRISP-DM are explained. Mentions Team Data Science Process TDS, launched in 2016 by Microsoft. They all might not handle the nuances of data science. Brechner (2015) referred to as Kanban. Akred (2016)

					Discussions on Agile for Data Science.
(Saltz et al., 2018)	•CRISP-CM •Team Data Science Process (TDSP) •Knowledge Discovery in Data Science (KDDS)	•Scrum •Kanban •Hybrid	Perhaps part of the reason that many teams use an ad hoc approach is that there is no clear agile process that is designed for data science projects.		This article provides many links between Scrum, Kanban CRIPS-DM for the Data Science domain. Saltz (2017): collection, preparation, visualization, management and preservation. Having no process in place: - thwarting efficiency / improvement team - slow information sharing - delivering the wrong thing - lack of reproducibility - poor coordination - scope creep Scrum, Kanban and CRISP-DM are explained. Mentions Team Data Science Process TDS, launched in 2016 by Microsoft. They all might not handle the nuances of data science. Brechner (2015) referred to as Kanban. Akred (2016) Discussions on Agile for Data Science.
(Sharma et al., 2012)	•CRISP-DM •Integrated Knowledge Discovery and Data Mining (IKDDM)		•perceived ease of use •perceived usefulness •semantic quality •user satisfaction		IKDDM outperforms CRISP-DM on -perceived ease of use -user satisfaction -perceived usefulness -perceived semantic quality

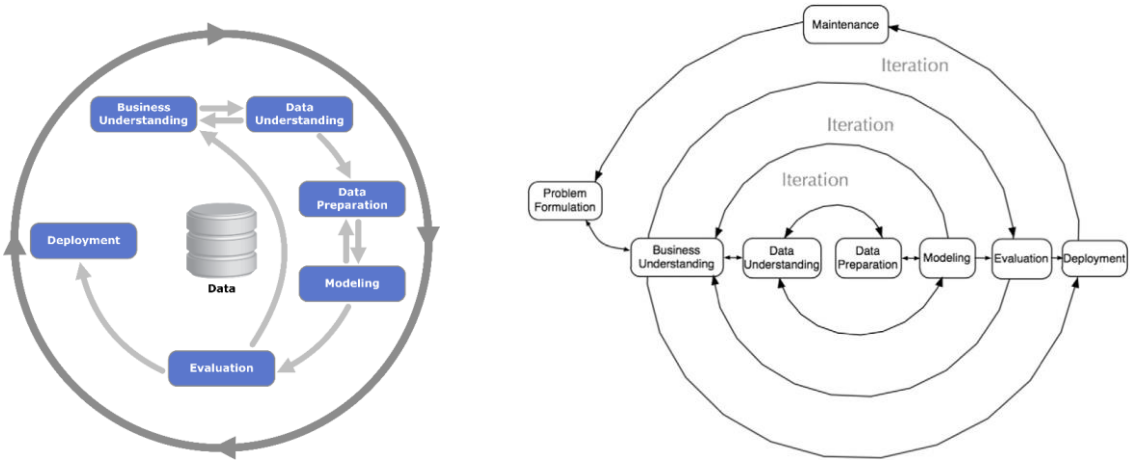
Appendix E DAP and DAM classes

Table 5 shows the various concepts found in the publications. During the analysis, it was found useful to split the DAP concept into project related and project context related. Not all articles discussed project context concepts.

	DAP	**DAP Context**	**DAM**
(Ahangama & Poo, 2015)	process management	Communication, management and knowledge.	
(Baijens & Helms, 2019)			Tasks, steps and practices
(Batra, 2018)	project technological capability, development method		Plan-driven, agile-driven, agile-plan-balanced
(Gao, Koronios, & Selle, 2015)	people, technology: identifiable business value, innovative analysis tools, adequate hardware, analytical skillset,	Information strategy	business, measurement, data, learning, analysis and implementation
(Jensen, Nielsen, & Persson, 2019)	business case, context, benefits, measures, relationships, missing benefits and end-users		
(Li, Thomas, & Osei-Bryson, 2016)	hypothesis generation, hypothesis testing	Enterprise Data Architecture	problem formulation, business understanding, data understanding, data preparation, modelling, evaluation, deployment and maintenance
(Mariscal, Marbán, & Fernández, 2010)			analysis, development and maintenance. Plus 17 subprocesses.
(Saltz, Shamshurin, & Connors, 2017)	infrastructure (computing needs) and discovery (clarity of questions). well-defined, hard-to-justify, exploratory and smaller-data.		
(Saltz, Hotz, Wild, & Stirling, 2018)	SCRUM, Kanban, hybrid: - thwarting efficiency / improvement team - slow information sharing - delivering the wrong thing - lack of reproducibility - poor coordination - scope creep		SCRUM, Kanban, hybrid

Table 10: DAP and DAM classes

Appendix F CRISP-DM and Snail Shell



Left: The Cross Industry Standard Process for Data Mining (CRISP-DM) model (Chapman et al., 2000).

Right: Snail Shell (Li et al., 2016).

Appendix G Questions

Interview questions:

- How would you describe the DAPs that you have or had at hand, what are the characteristics?
- What can you tell about the hypothesis testing, smaller/big data, type of requirements and organizational data governance in your DAPs? – The DAP model was used by to guide the conversation without showing it.
- Are you aware of any specific DAMs for DAPs? If so, which DAMs do you use?
- Which process steps do you distinguish in your projects?

--showing the DAM dimension model --

- What would be your feedback on the DAM model?
- How can you map your DAPs to the DAM classes and why is that?
- Do you have final reflections, eventually also in term of current trends?

Focus group questions:

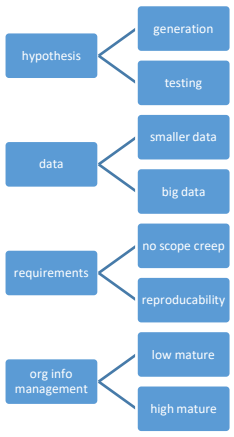
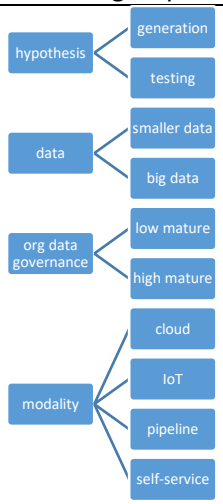
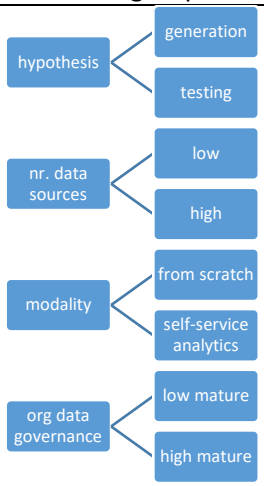
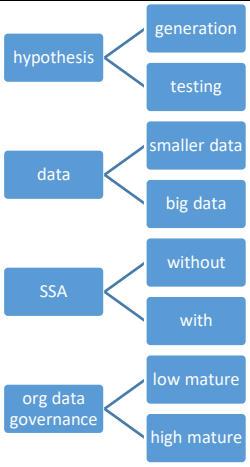
Do you have feedback on:

- The DAP dimension model?
- The presented mappings?
- Could you please suggest another mapping (s) which you believe are relevant to consider?

Appendix H Coding scheme

First level code	Second level code	Third level code
hypothesis	hypothesis generation	
	hypothesis testing	
data	smaller data	structured data
	big data	
requirements	scope creep	
	reproducibility	
data	data governance	maturity
	information management	
	data architecture	
orientation	waterfall	
	agile	
process step	analysis	problem formulation
		business understanding
		data understanding
	development	data preparation
		modelling
	maintenance	evaluation
		deployment

Appendix I: DSR iterations

Interviews	Focus group 1	Focus group 2	Final
			

Modality: Based on the interviews, a new DAP class was introduced called modality with the subclasses *cloud*, *IoT*, *pipeline* and *self-service*. Only self-service was found to be useful by the focus groups, the others were not found to be relevant in the context of DAM. Concepts like IoT and pipeline were considered as the toolset of the project. The concept of cloud was also considered, even though during the interviews many respondents referred to it as a fundamental change. In the second focus group the modality class was tested with the subclasses *from scratch* and *self-service*, which was later slightly changed into with and without.

Appendix J: Pipeline and Continuous Integration

Queries performed with Web of Science on May 31, 2020.

Query	Nr of results on topic	Nr of results on Title
pipeline "machine learning"	1052	37
pipeline "predictive analytics"	7	0
pipeline "descriptive analytics"	1	0
pipeline "business intelligence"	10	0
pipeline "data analytics"	66	5

Query	Number of results on topic	Number of results on Title
"continuous integration" "machine learning"	7	0
"continuous integration" "predictive analytics"	1	1
"continuous integration" "descriptive analytics"	0	0
"continuous integration" "business intelligence"	0	0
"continuous integration" "data analytics"	0	0